Supplementary Material

A Review: Non-linear ICA via time contrastive learning

Here we provide a more detailed review of the theory of TCL. For further details we refer readers to Hyvärinen and Morioka (2016) and we note that the results presented in this section are adapted from Hyvärinen and Morioka (2016). We begin by showing that an optimally discriminative feature extractor combined with a linear multinomial classification layer learns to model the non-stationary probability density of observations within each experimental condition.

Recall that we observe d-dimensional data, $\mathbf{X}(i) = \mathbf{f}(\mathbf{S}(i))$, which is generated via a smooth and invertible non-linear mixture, $\mathbf{f}$, of d independent latent variables, $\mathbf{S}(i)$. As in linear ICA, the latent variables are assumed to be mutually independent. However, we also assume they are non-stationary. In particular, we assume the distribution of latent variables to be piece-wises stationary such that we may associate a label, $C_i \in \mathcal{E}$, with each $\mathbf{S}(i)$ indicating the piece-wise stationary segment from which it was generated. In this manner, it is assumed that the distribution of latent variables varies across segments, as shown in Figure S.1. As such, we write $C_i \in \mathcal{E}$ to denote the segment of the ith observation where $\mathcal{E} = \{1, \ldots, E\}$ is the set of all distinct segments. For example, each segment may correspond to a distinct experimental condition. As the function $\mathbf{f}$ is smooth and invertible, it follows that the distribution of each $\mathbf{X}(i)$ will also vary across segments.

We may therefore consider the task of classifying observed data into the various segments as a multinomial classification task consisting of features, $\mathbf{X}(i)$, and categorical labels, C_i. For any observation, $\mathbf{X}(i)$, associated with true label $C_i \in \mathcal{E}$, we have:

$$p(C_i = \tau | \mathbf{X}(i), \theta, \mathbf{W}, \mathbf{b}) = \frac{\exp\left(\mathbf{w}_\tau^T \, \mathbf{h}(\mathbf{X}(i); \theta) \; + b_\tau\right)}{1 + \sum_{e=2}^{E} \exp\left(\mathbf{w}_e^T \, \mathbf{h}(\mathbf{X}(i); \theta) \; + b_e\right)}, \tag{S.1}$$

where θ are parameters for the neural network feature extractor and the weight matrix, $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_E]$, and bias vector, $\mathbf{b} = [b_1, \ldots, b_E]$, parameterize the final multinomial layer. We note that the sum in the denominator goes from $e = 2, \ldots, E$. This is because we fix $\mathbf{w}_1 = 0$ and $b_1 = 0$ in order to avoid indeterminancy in the softmax function.

Conversely, we can derive the true posterior distribution over the label C_i as:

$$p(C_i = \tau | \mathbf{X}(i)) = \frac{p(\mathbf{X}(i) | C_i = \tau) \; p(C_i = \tau)}{\sum_{e=1}^{E} \; p(\mathbf{X}(i) | C_i = e) \; p(C_i = e)}. \tag{S.2}$$

If we assume that the feature extractor has a universal function approximation capacity and that we have infinite data then the multinomial logistic classifier based on features $\mathbf{h}(\mathbf{X}; \theta)$ will converge to the optimal classifier, implying that equation (S.1) will equal equation (S.2) for all $\tau \in \mathcal{E}$. We may then consider the following ratio:

$$\frac{p(C_i = \tau | \mathbf{X}(i), \theta, \mathbf{W}, b)}{p(C_i = 1 | \mathbf{X}(i), \theta, \mathbf{W}, b)} = \frac{p(C_i = \tau | \mathbf{X}(i))}{p(C_i = 1 | \mathbf{X}(i))}, \tag{S.3}$$

which after expanding and taking logarithms yields:

$$\mathbf{w}_\tau^T \; \mathbf{h}(\mathbf{X}(i); \theta) + b_\tau = \log \; p(\mathbf{X}(i) | C_i = \tau) - \log \; p(\mathbf{X}(i) | C_i = 1) + \log \; \frac{p(C_i = \tau)}{p(C_i = 1)}, \tag{S.4}$$

indicating that the optimal feature extractor computes the log probability density function of the data within each experimental condition (relative to some pivot segment, in this case the first condition). We note that this condition holds for all $\tau \in \mathcal{E}$.

If we further assume the data were generated according to a non-stationary ICA model as described in equation (3), then equation (S.4) yields:

$$W_\tau^T \; \mathbf{h}(\mathbf{X}; \theta) - k_1(\mathbf{X}(i)) = \sum_{j=1}^{d} \; \lambda_j(\tau) \; q(S_j) - k_2(\tau), \tag{S.5}$$

where the sum is taken over each independent source, $S_1, \ldots, S_d$. Equation (S.5) follows from the change of variable from $\mathbf{S}$ to $\mathbf{X}$, noting that the Jacobians required by such a transformation cancel out because of the subtraction in the

right hand side of equation (S.4). As a result, by modeling the log probability densities with respect to some pivot segment (in this case segment 1), we do not need explicity compute the Jacobians. We note that k_1 is a function which does not depend on τ and k_2 is a function which does not depend on either $\mathbf{X}$ or $\mathbf{S}$. As a result, it follows that both $\mathbf{h}(\mathbf{X}; \theta)$ and $q(\mathbf{S})$ must span the same linear space, implying that we may compute latent sources up to some non-linearity, $q(\mathbf{S})$, by first learning a feature extractor based on TCL and subsequently applying linear ICA on estimated features, $\mathbf{h}(\mathbf{X}; \theta)$. Figure S.1 summarizes the relationship between the generative model and TCL.

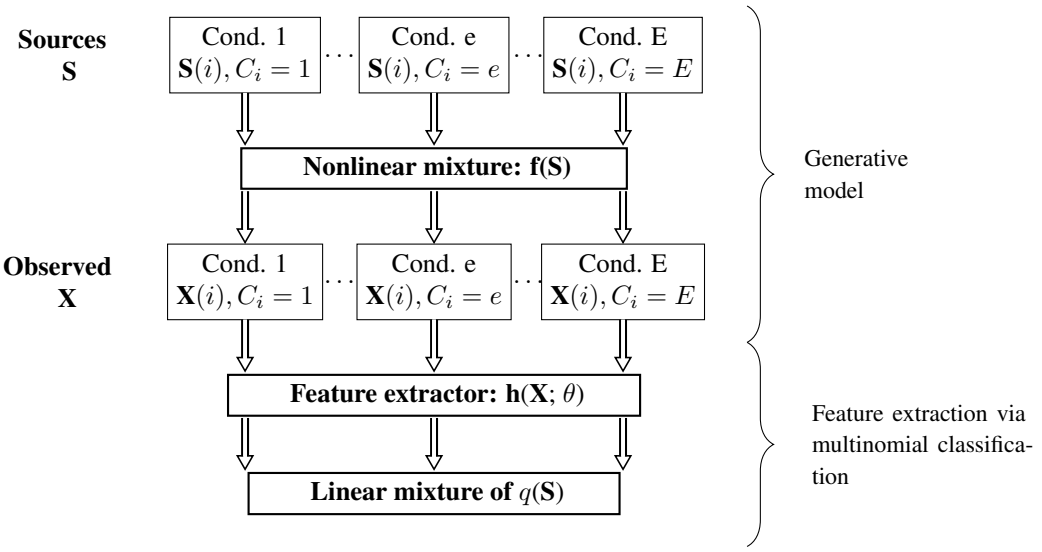

Figure S.1: A cartoon visualization which highlights the relationship between features learnt using TCL and a non-linear ICA model. The non-linear ICA model assumes observations, $\mathbf{X}$, are generated based on non-stationary latent sources whose distribution varies according the distinct experimental conditions, $e \in \mathcal{E}$.

B Relationship between ICA and SEM

In this section we formally outline the relationship between a bivariate non-linear ICA model and a non-linear SEM, as first discussed in Section 3.1. Recall, under a non-linear ICA model we observe data, $\mathbf{X}$, which is generated according to a non-linear mixture of independent latent variables, $\mathbf{S}$, such that:

$$\mathbf{X} = \mathbf{f}(\mathbf{S}). \tag{S.6}$$

Conversely, a SEM is defined as a collection of structural equations which explicitly define the generative mechanisms for each observed variable as a function of its parents in the causal graph and its latent disturbance. In the context of bivariate data, and assuming that $X_1 \to X_2$, the structural equations may be written as:

$$X_1 = f_1(N_1), \tag{S.7}$$
$$X_2 = f_2(X_1, N_2). \tag{S.8}$$

The correspondence between ICA and SEMs was first noted in the linear case by Shimizu et al. (2006). In such a setting we have that:

$$\mathbf{X} = \mathbf{A}\mathbf{S}, \tag{S.9}$$

where $\mathbf{A} \in \mathbb{R}^{2\times 2}$ denotes the ICA mixing matrix. In the linear setting, there is an exact correspondence between the structural equations, f_1 and f_2, and the linear mapping from sources to observations parameterized by $\mathbf{A}$. Formally, each structural equation will correspond to a row in the ICA mixing matrix. This can be seen by expanding equation (S.9) to yield:

$$X_1 = a_{1,1}S_1 + a_{1,2}S_2, \tag{S.10}$$
$$X_2 = a_{2,1}S_1 + a_{2,2}S_2, \tag{S.11}$$

where $a_{l,k}$ denotes the (l,k) entry of $\mathbf{A}$. Note that under the assumption that $X_1 \to X_2$ we have that $a_{1,2} = 0$ and we may re-write equation (S.11) as:

$$X_2 = \frac{a_{2,1}}{a_{1,1}} X_1 + a_{2,2} S_2. \tag{S.12}$$

This serves to highlight the correspondence between equation (S.6) and equations (S.7-S.8). Equation (S.12) also serves to demonstrate the fact that while there is a correspondence between the ICA model and the structural equations, these will not necessarily be equal, as some transformations have been applied to coefficients in equation (S.12). This effect will occur to a larger extent in the context of non-linear ICA and non-linear SEMs as instead of transformation by a scalar there will be an arbitrary non-linear transformation. Finally, each latent disturbance, N_j, will correspond to a latent source, $S_{\pi(j)}$.

C A linear ICA method for piece-wise stationary sources

Formally, assumptions 1-3 of Theorem 1 guarantee that TCL, as presented in Section 2.2, will recover a linear mixture of latent independent sources up to point-wise transformation. This implies that the hidden representations obtained satisfy:

$$\mathbf{h}(\mathbf{X}; \theta) = \mathbf{A} q(\mathbf{N}), \tag{S.13}$$

for some linear mixing matrix, $\mathbf{A}$. Equation (S.13) suggests that applying ordinary linear ICA to hidden representations, $\mathbf{h}(\mathbf{X}; \theta)$, will allow us to recover $q(\mathbf{N})$. However, the use of linear ICA is premised on the assumption that latent variables are independent. This is not necessarily guaranteed under the generative model presented in equation (3) in the context of a fixed number of segments. While it is certainly true that N_1 and N_2 are conditionally independent given segment labels, it may be the case that they are not marginally independent over all segments; for example it is possible that their variances are somehow related (e.g., they may be monotonically increasing). We note that this is not an issue when the number of segments grows asymptotically and we assume exponential family parameters, $\lambda_j(e)$, are randomly generated, as stated in Corollary 1 of Hyvärinen and Morioka (2016).

Here we seek to address this issue by proposing an alternative linear ICA algorithm which explicitly models the piece-wise stationary nature of the data. In particular, the proposed linear ICA algorithm explicitly incorporates equation (3) as the generative model for latent sources. As such, it can be used to accurately unmix sources in the final stage of TCL, especially when the number of segments in small.

To set notation, we assume we observe data which corresponds to a linear mixture of sources:

$$\mathbf{Z} = \mathbf{A}\mathbf{S}$$

where $\mathbf{Z}, \mathbf{S} \in \mathbb{R}^d$ and $\mathbf{A} \in \mathbb{R}^{d \times d}$ is a square mixing matrix. Popular ICA algorithms, such as FastICA, estimate the unmixing matrix $\mathbf{W} = \mathbf{A}^{-1}$ by maximizing the log-likelihood under the assumptions that sources are independent and non-Gaussian. The objective function for FastICA is therefore of the form:

$$\log p(\mathbf{Z}) = \sum_{j=1}^{d} q(\mathbf{w}_j^T \mathbf{Z}) - Z(W), \tag{S.14}$$

where $Z(W) = \log |\det W|$ is the normalization constant and we write $\mathbf{w}_k$ to denote the kth row of $\mathbf{W}$. Typically a parametric form is assumed for $q(\cdot)$, popular examples being exponential or log-cosh. In the context of the generative model for sources specified in equation (3), the FastICA algorithm therefore proceeds under the assumption that $\lambda_j(e)$ is constant across all segments $e \in \mathcal{E}$.

In order to address this issue we consider an alternative model for the density of latent sources. In particular, we seek to directly employ the piece-wise stationary log-density described in equation (3). As such, we log-density of an observation within segment $e \in \mathcal{E}$ is defined as:

$$\log p_e(\mathbf{Z}) = \sum_{j=1}^{d} \lambda_j(e) q_j(\mathbf{w}_j^T \mathbf{Z}) - Z(W, \lambda(e)). \tag{S.15}$$

In contrast to equation (S.14), the log-density of each observation depends on both the segment, e, the exponential family parameters, $\boldsymbol{\lambda} = \{\lambda_j(e) : e \in \mathcal{E}, j = 1, \ldots, d\}$, as well as the unmixing matrix, $\mathbf{W}$.

In order to recover latent sources we propose to estimate parameters, corresponding to unmixing matrix as well as exponential family parameters, via score matching (Hyvärinen, 2005). This avoids the need to estimate the normalization parameter, which may not be available analytically when sources follow unnormalized distributions. The score matching objective for the ICA model defined in equation (S.15) is defined as:

$$\begin{aligned}\tilde{J} = &\sum_{e\in\mathcal{E}}\sum_{j=1}^{d}\lambda_j(e)\frac{1}{n_e}\sum_{i,C_i=e} q_j''(\mathbf{w}_j^T\mathbf{Z}(i)) \\ &+\frac{1}{2}\sum_{e\in\mathcal{E}}\sum_{j,k=1}^{d}\lambda_k(e)\lambda_j(e)\mathbf{w}_k^T\mathbf{w}_j\frac{1}{n_e}\sum_{i,C_i=e} q_k'(\mathbf{w}_k^T\mathbf{Z}(i))q_j'(\mathbf{w}_j^T\mathbf{Z}(i)),\end{aligned} \tag{S.16}$$

where we write q_j' and q_j'' to denote the first and second derivatives of q_j with respect to observations, $\mathbf{Z}$. We propose to minimize equation (S.16) via block gradient descent, conditionally updating the mixing matrix $\mathbf{W}$ and exponential family parameters $\boldsymbol{\lambda}$. This has the important benefit that conditional on $\mathbf{W}$, there is a closed form update for $\boldsymbol{\lambda}$ (Hyvärinen, 2007).

Experimental results

In order to assess the performance of the proposed linear ICA algorithm, we generated bivariate data following the piece-wise stationary distribution described in equation (3). We compare the performance of the proposal algorithm against the following popular linear ICA algorithms: FastICA (Hyvärinen, 1999), Infomax ICA (Bell and Sejnowski, 1995) and Joint Diagonalization method proposed by Pham and Cardoso (2001) which also accommodates non-stationary sources.

In order to assess the performance of the proposed method we consider two scenarios:

- The exponential family parameters are deliberately generated such that there is a statistical dependence structure across segments. In particular, we generate bivariate data where we explicitly enforce $\lambda_j(e)$to be monotonically increasing in e. As a concrete example, when sources follow a Laplace distribution this implies that $q(S) = |S|$ and in turn $\lambda_j(e)$ corresponds to the variance of the jth source in segment e. In such a setting, we generate piece-wise stationary Laplace sources with where the variances are correlated across segments.
- As a baseline, we also generate data where the exponential family parameters are generated at random. This removes any systematic, higher-order dependence between latent sources.

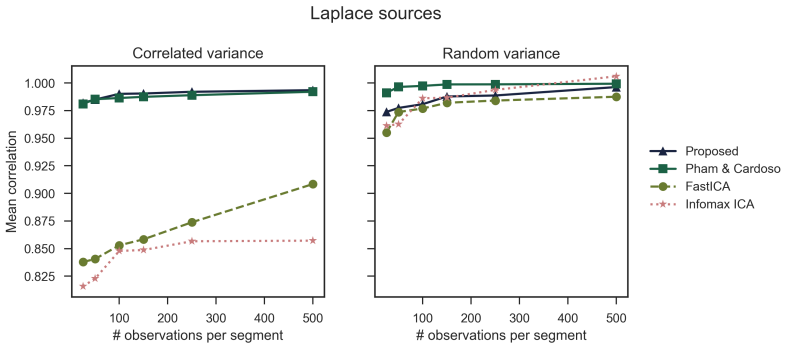

Figure S.2: Performance of various linear ICA algorithms for piece-wise stationary Laplace sources. The left panel shows performance when the variance of latent sources are positive correlated, introducing second order dependence in the data. As expected, FastICA and Infomax ICA perform poorly in this context. The right panel shows similar data were variances are no longer correlated, resulting in good performance for all algorithms.

We begin by generating bivariate data were sources follow a piece-wise stationary Laplace distribution. This implies that sources follow the log-density specified in equation (3) where $q_j(S_j) = |S_j|$ and each term $\lambda_j(e)$ denotes the

variance of the jth source in segment e. Results when data is generated over five segments are provided in Figure S.2. The left panel shows the case were the variances of each latent source are correlated across segments. We note that when this is the case, methods such FastICA and Infomax ICA perform poorly. This is in contrast to the proposed method and the joint diagonalization approach of Pham and Cardoso (2001), who explicitly model the non-stationary nature of the data. The right panel of Figure S.2 shows equivalent results when variances are randomly generated, thereby removing second order dependence between latent variables. As expected, in this setting all methods perform well.

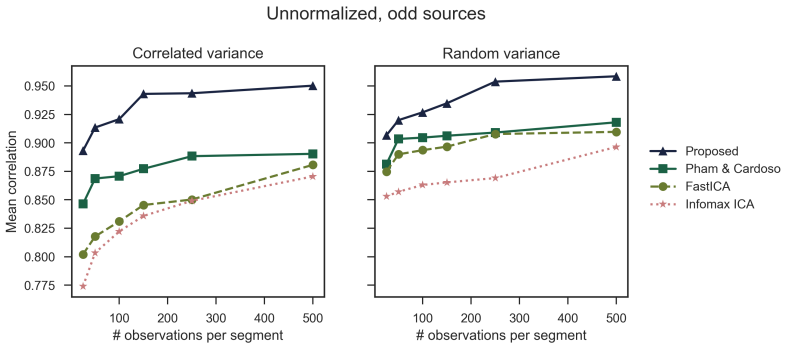

Figure S.3: Performance of various linear ICA algorithms when latent sources follow the log-density detailed in equation (S.17). The left panel shows performance when exponential family parameters, $\lambda_i(e)$, are positive correlated, introducing higher order dependence in the data. The right panel shows similar data were parameters, $\lambda_i(e)$ are randomly generated.

In order to further probe the differences between the proposed method and the approach of Pham and Cardoso (2001), we consider latent sources with an unnormalized distribution. In particular, we generate sources such that the log-density within a particular segment is as follows:

$$\log p_e(S_j) = \begin{cases} -3\lambda_j(e)|S_j| - \frac{1}{2}S_j^2, & \text{if } S_j \geq 0 \\ -\lambda_j(e)|S_j| - \frac{1}{2}S_j^2, & \text{otherwise.} \end{cases} \tag{S.17}$$

Such a density is both unnormalized but also skewed. As such, we expect the joint diagonalization algorithm of Pham and Cardoso (2001) to perform poorly in this setting as it exclusively studies covariance structure and therefore cannot model skewed distributions. Figure S.3 visualizes the results for these experiments. As expected, the Joint Diagonalization algorithm of Pham and Cardoso (2001) suffers a drop in performance. However, it continues to outperform FastICA and Infomax ICA, especially when there are dependencies over the exponential family parameters. We note that the proposed model, where parameters are estimated using score matching, is shown to be more robust.

D Proof of Theorem 2

The proof of Theorem 2 follows from a combination of the presented assumptions together with Property 1. Formally, assumptions 1–3 guarantee that the TCL, as presented in Section 2.2, will recover a linear mixture of latent independent sources up to point-wise transformation. This, combined with the novel linear ICA algorithm described in Section 3.2, imply that, in the limit of infinite data (as stipulated in Theorem 1 of Hyvärinen and Morioka (2016)), latent disturbances can be recovered. In particular, we note Assumption 2 states that segments must have distinct distributions, thus implying that the matrix $\mathbf{L}$ with elements $\mathbf{L}_{e,j} = \lambda_j(e) - \lambda_j(1)$ for $e = 1, \ldots, E$ and $j = 1, \ldots, d$ will have full rank.

We note that in practice we recover the latent disturbances up to element-wise transformation, $q(\mathbf{N})$, as opposed to $\mathbf{N}$. This is not a problem when a general test of statistical independence, such as HSIC which can capture arbitrary (i.e., non-linear) dependencies, is employed. Finally, Assumption 1 further states there is no latent confounder is present, implying that running all possible pairwise tests using a sufficiently flexible independence test, as required by assumption 4, will allow us to determine causal structure.

E Relationship between likelihood ratio and measures of independence

In this section we derive the result presented in equation (8), for some permutation π of latent disturbances. We begin by considering the mutual information between X_1 and $N_{\pi(2)}$:

$$\begin{aligned} I(X_1, N_{\pi(2)}) &= H(X_1) + H(N_{\pi(2)} - H(X_1, N_{\pi(2)}) \\ &= H(X_1) + H(N_{\pi(2)} - H(X_1, X_2) - \mathbb{E}\left[\log \left| \frac{\partial \mathbf{g}_{\pi(2)}}{\partial X_2} \right| \right] \end{aligned}$$

where we employ the same change of variable, whose Jacobian can be easily evaluated, as in Section 3.4. In particular, we have used the property that:

$$H(X_1, N_{\pi(2)}) = H(X_1, X_2) + \log |\det \mathbf{J}\tilde{\mathbf{g}}|$$

and noted that the particular choice of $\tilde{\mathbf{g}}$ allows us to directly compute the Jacobian as $\frac{\partial \mathbf{g}_{\pi(2)}}{\partial X_2}$. We may therefore compute the difference in mutual information between each observed variable and the relevant latent disturbance, yielding:

$$\begin{aligned} I(X_1, N_{\pi(2)}) - I(X_2, N_{\pi(1)}) = H(X_1) + H(N_{\pi(2)} - \mathbb{E}\left[\log \left| \frac{\partial \mathbf{g}_{\pi(2)}}{\partial X_2} \right| \right] \\ - H(X_2) - H(N_{\pi(1)} + \mathbb{E}\left[\log \left| \frac{\partial \mathbf{g}_{\pi(1)}}{\partial X_1} \right| \right] \end{aligned}$$

which is precisely the negative of likelihood ratio presented in Section 3.4.

F Baseline methods

In this section we briefly overview and provide pseudo-code for alternative methods which are presented as baselines in the manuscript.

- **DirectLiNGAM**: The DirectLiNGAM method of Shimizu et al. (2011) is based on the property that within a linear non-Gaussian acyclic model (LiNGAM), if we regress out the parents of any variable, then the residuals also follow a LiNGAM. Based on this property, the authors propose an iterative algorithm through which to iteratively uncover exogenous variables. oFurther, if variables follow a LiNGAM, then due to the additive nature of noise in such models, the residuals will be independent when we regress the parent on its children. As a result, we may infer the causal structure by studying the statistical independences between variables and residuals. Pseudo-code for the bivariate DirectLiNGAM method is provided in Algorithm 1.

Algorithm 1: Bivariate causal discovery using DirectLiNGAM (Shimizu et al., 2011)

Input : Bivariate data, $\mathbf{X}$, and significance level α.

for $i \in \{1, 2\}$ **do**

  Linearly regress X_i on $X_{\{1,2\}\setminus i}$ and compute the residual $\hat{N}_i$.

  Evaluate the test:

$$H_{X_i,\hat{N}_i,0} : \mathbf{P}_{X_i,\hat{N}_j} = \mathbf{P}_{X_i}\mathbf{P}_{\hat{N}_j} \text{ against } H_{X_i,\hat{N}_j,1} : \mathbf{P}_{X_i,N_j} \neq \mathbf{P}_{X_i}\mathbf{P}_{\hat{N}_j} \text{ at the } \tfrac{\alpha}{2} \text{ level.}$$

end

if *we fail to reject the null hypothesis only once* **then**

  Variable i' such that we fail to reject $H_{X_{i'},\hat{N}_j,0}$ is the cause variable

else

  The causal dependence structure is inconclusive

end

- **RESIT**: The RESIT method, first proposed by Hoyer et al. (2009) and subsequently extended by Peters et al. (2014), can be seen as a non-linear extension of the DirectLiNGAM algorithm. The RESIT algorithm is premised on the assumption of an additive noise model (ANM), which implies that each structural equation are of the form $X_j = f_j(\mathbf{PA}_j) + N_j$. Given the ANM assumption, RESIT is able to recover the causal structure by testing for dependence between variables and residuals. Gaussian process regression is employed in order to accommodate for non-linear additive causal relations.

Algorithm 2: Bivariate causal discovery using RESIT (Hoyer et al., 2009; Peters et al., 2014)

Input : Bivariate data, $\mathbf{X}$, and significance level α.

for $i \in \{1, 2\}$ **do**

  Regress X_i on $X_{\{1,2\}\setminus i}$ and compute the residual $\hat{N}_i$. `// Gaussian process regression is employed`

  Evaluate the test:

$$H_{X_i,\hat{N}_i,0} : \mathbf{P}_{X_i,\hat{N}_j} = \mathbf{P}_{X_i}\mathbf{P}_{\hat{N}_j} \text{ against } H_{X_i,\hat{N}_j,1} : \mathbf{P}_{X_i,N_j} \neq \mathbf{P}_{X_i}\mathbf{P}_{\hat{N}_j} \text{ at the } \tfrac{\alpha}{2} \text{ level.}$$

end

if *We fail to reject the null hypothesis only once* **then**

  Variable i' such that we fail to reject $H_{X_{i'},\hat{N}_j,0}$ is the cause variable

else

  The causal dependence structure is inconclusive

end

- **Non-linear ICP**: The ICP method proposed by Peters et al. (2016) proposes a fundamentally different approach to causal discovery. The underlying principle of the ICP algorithm is that the direct causal predictors of a given variable must remain invariant across distribution shifts induced by various experimental conditions. In the context of bivariate data, the non-linear ICP algorithm, as described in Section 6.1 of Peters et al. (2016) therefore corresponds to fitting a non-linear regression model on the data across all experimental conditions and testing whether the distribution of residuals is the same within each condition. We note that such an approach assumes an additive noise model, as this greatly simplifies testing for invariance.

Algorithm 3: Bivariate causal discovery via non-linear ICP (Peters et al., 2016)

Input : Bivariate data, $\mathbf{X}(i)$, labels $C_i \in \{1, \ldots, E\}$ and significance level α.

for $i \in \{1, 2\}$ **do**
  Regress X_i on $X_{\{1,2\}\setminus i}$ and compute the residual, $\hat{N}_i$
  `// Note that Gaussian process regression is employed and we aggregate data across all experimental conditions`
  Evaluate the test:

$$H_{\hat{N}_i,0} : \mathbf{P}_{\hat{N}_i,e} = \mathbf{P}_N \text{ for all } e \in \mathcal{E} \text{ against } H_{\hat{N}_i,1} : \mathbf{P}_{\hat{N}_i,e} \neq \mathbf{P}_N \text{ for some } e \in \mathcal{E}$$

  where $\mathbf{P}_N$ is some arbitrary distribution. `// The Kolmogorov-Smirnov test is employed`
end
if *We fail to reject the null hyp. only once* **then**
  Variable i such that we fail to reject $H_{\hat{N}_i,0}$ is the cause variable
else
  The causal dependence structure is inconclusive
end

- **RECI**: Blöbaum et al. (2018) propose a method for inferring the causal relation between two variables by comparing the regression errors in each possible causal direction. Under some mild assumptions, they are able to prove that the magnitude of residual errors should be smaller in the causal direction. This suggests a straightforward causal discovery algorithm, which we outline below. We note that while any non-linear regression method may be employed, our implementation used Gaussian process regression.

Algorithm 4: Bivariate causal discovery using RECI (Blöbaum et al., 2018)

Input : Bivariate data, $\mathbf{X}$.

Standardize data such that $\mathbf{X}_t$ is zero mean and unit variance.
for $i \in \{1, 2\}$ **do**
  Regress X_i on $X_{\{1,2\}\setminus i}$ and evaluate the mean-squared error, MSE_i `// Gaussian process regression is employed`
end
if $MSE_1 < MSE_2$ **then**
  Variable 1 is the cause variable
else
  Variable 2 is the cause variable
end

- **CD-NOD**: Zhang et al. (2017) propose a causal discovery algorithm which explicitly accounts for non-stationarity or heterogeneity over observed variables. The CD-NOD algorithm accounts for non-stationarity, which may manifest itself as changes in the causal modules, by introducing an additional variable representing the time or domain index into the causal DAG. Conditional independence testing is then employed to recover the skeleton over the augmented DAG. Their method can find causal direction by making use of not only invariance, but also independent changes of causal models, as an extended notion of invariance. We also note that CD-NOD is a non-parametric method, implying it is able to accommodate non-linear causal dependencies.

Algorithm 5: Bivariate causal discovery using CD-NOD (Zhang et al., 2017)

Input : Bivariate data, $\mathbf{X}$.

Build an augmented dataset consisting of $\mathbf{X}$ and C, the observed variable representing time or domain index.

`// Detection of changing modules`

for $i \in \{1, 2\}$ **do**

  Test for marginal and conditional dependence between X_i and C

  If they are conditionally independent given $X_{\{1,2\}\setminus i}$ then we remove the edge between X_i and C in the augmented DAG

end

`// Recover causal skeleton`

if $X_1 \perp\!\!\!\perp X_2 \mid C$ **then**

  Remove the edge between X_1 and $X_2 \Rightarrow$ no causal relation between X_1 and X_2

else

  if *Only one of* X_1 *and* X_2 *is marginally or conditionally dependent on* C **then**

    Dependent variable is reported as the cause

  else

    Determine cause variable by comparing mutual information

  end

end

G Pseudo-code for NonSENS

In this supplementary we provide pseudo-code for the proposed NonSENS method, as described in Section 3.3.

Algorithm 6: Bivariate causal discovery via NonSENS

Input : Bivariate time-series data, $\mathbf{X}(i)$, labels $C_i \in \mathcal{E}$ and significance level α.

Estimate $(\hat{N}_1(i), \hat{N}_2(i))$ via TCL where the final linear unmixing is performed using the ICA algorithm presented in Section 3.2.

for $i \in \{1, 2\}$ *and* $j \in \{1, 2\}$ **do**

  Evaluate the test:

$$H_{X_i,\hat{N}_j,0} : \mathbf{P}_{X_i,\hat{N}_j} = \mathbf{P}_{X_i}\mathbf{P}_{\hat{N}_j} \text{ against } H_{X_i,\hat{N}_j,1} : \mathbf{P}_{X_i,N_j} \neq \mathbf{P}_{X_i}\mathbf{P}_{\hat{N}_j} \text{ at the } \tfrac{\alpha}{4} \text{ level.}$$

end

if *We fail to reject the null hyp. only once* **then**

  Variable i' such that we fail to reject $H_{X_{i'},\hat{N}_j,0}$ for some $j \in \{1, 2\}$ is the causal variable

else

  The causal dependence structure is inconclusive

end

H Further experimental results

In this section of the supplementary material we present further experimental results. In particular, in Section H.1 we present results for bivariate causal discovery in the context of a fixed number of experimental conditions, $|\mathcal{E}| = 10$, and increasing observations per segment. In Section H.2 we provide addition results in the context of multivariate causal discovery. In particular, we report the Hamming distance between true and estimated DAGs.

H.1 Additional bivariate causal discovery experiments

We consider the performance of all algorithms in the context of a fixed number of experimental conditions, $|\mathcal{E}| = 10$, and an increasing number of observations per condition, n_e. The results are presented in Figure S.4, where we repeat each experiment 100 times. We note that all algorithms are able to accurately identify causal structure in the presence of LiNGAMs (corresponding to a 1 layer mixing-DNN). However, as the causal structure becomes increasingly non-linear, the performance of all methods declines. In particular, we note that the proposed method has comparable performance with alternative methods such as RESIT and CD-NOD when the number of samples is small. However, as the number of observations increases the proposed method is able to out-perform alternative algorithms.

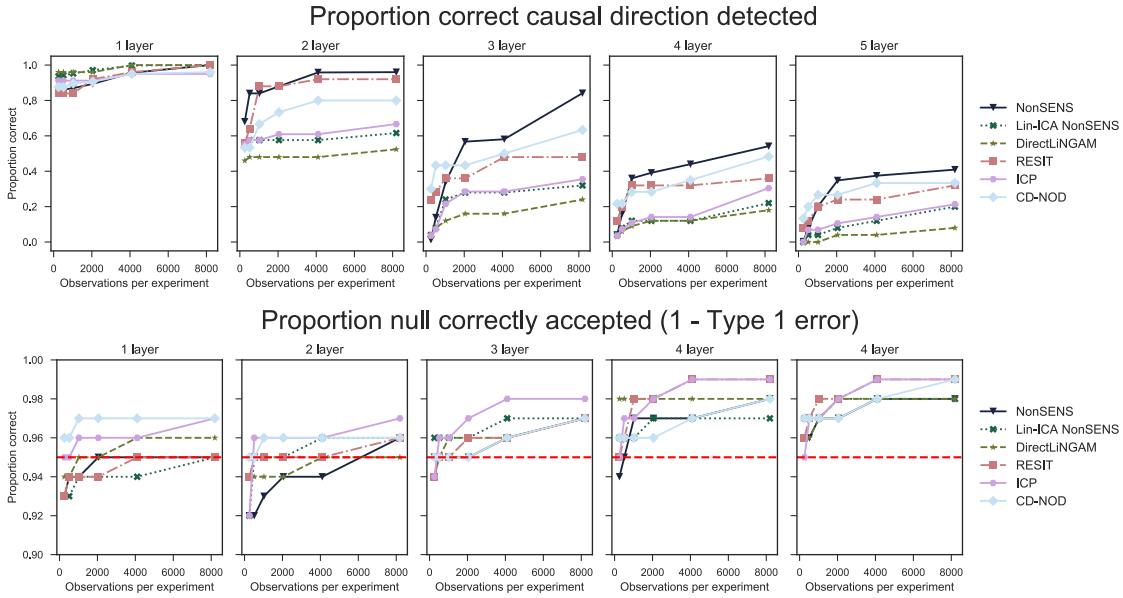

Figure S.4: Experimental results indicating performance as we increase the number of observations, n_e conditions, within each experimental condition for a fixed number of experimental conditions, $|\mathcal{E}| = 10$. Each horizontal plane plots results for varying depths of the mixing-DNN, ranging from $l = 1, \ldots, 5$. The top panel plots the proportion of times the correct cause variable is identified when a causal effect exists. The bottom panel considers data where no acyclic causal structure exists ($\mathbf{A}^{(l)}$ is not lower-triangular) and reports the proportion of times no causal effect is correctly reported. The dashed, horizontal red line indicates the theoretical $(1-\alpha)\%$ true negative rate. For clarity we omit the standard errors, but we note that they were small in magnitude (approximately $2-5\%$).

H.2 Multivariate causal discovery results

In this section we provide additional performance metrics in the context of multivariate causal discovery. While Figure 4 reported the F_1 score, we further provide results for the Hamming distance between true and estimated DAGs in Figure S.5.

.

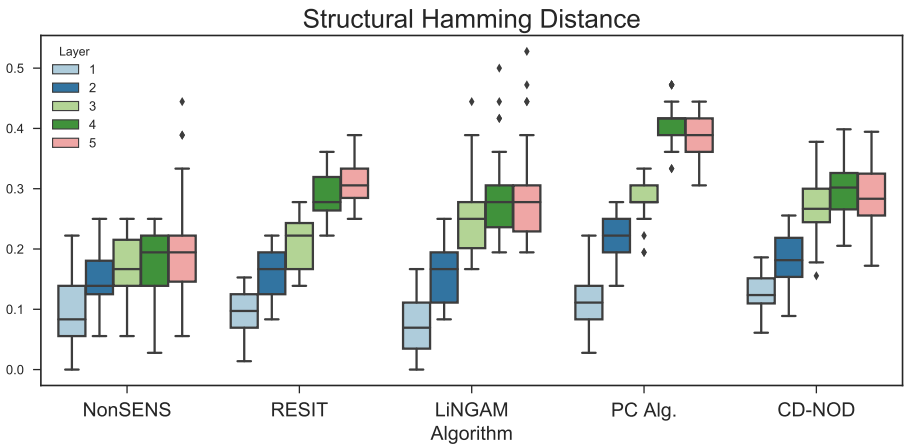

Figure S.5: Hamming distance results for multivariate causal discovery with 6-dimensional data. For each algorithm, we plot the structural Hamming distance as we vary the depth of the mixing-DNN from $l = 1, \ldots, 5$. Lower scores indicate better performance.

.

I Hippocampal functional MRI data

In this section we provide further details of the Hippocampal fMRI data employed in Section 4.2. The data was collected as part of the MyConnectome project, presented in Poldrack et al. (2015), which involved daily fMRI scans for a single individual (Caucasian male, aged 45). The data may be freely downloaded from `https://openneuro.org/datasets/ds000031/`.

We focus only on the resting-state fMRI data taken from this project, noting that future work may also wish to study the other modalities of data collected.

Data was collected from the same subject over a series of 84 successive days, allowing us to consider data collected on distinct days as a distinct experimental condition. Full details of the data acquisition pipelines are provided in Poldrack et al. (2015). For each day, we observe 518 BOLD measurements. After preprocessing, data was collected from the following brain regions: perirhinal cortex (PRC), parahippocampal cortex (PHC), entorhinal cortex (ERC), subiculum (Sub), CA1, and CA3/Dentate Gyrus (DG). This resulted in $d = 6$ dimensional data. As such, data employed consists of $n_e = 518$ observations per experimental condition and $|\mathcal{E}| = 84$ distinct conditions.